

# Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification

Xilun Chen<sup>†</sup>

xlchen@cs.cornell.edu

Yu Sun<sup>†</sup>

ys646@cornell.edu

Ben Athiwaratkun<sup>‡</sup>

pa338@cornell.edu

Claire Cardie<sup>†</sup>

cardie@cs.cornell.edu

Kilian Weinberger<sup>†</sup>

kqw4@cornell.edu

<sup>†</sup>Dept. of Computer Science, Cornell University, Ithaca NY, USA

<sup>‡</sup>Dept. of Statistical Science, Cornell University, Ithaca NY, USA

## Abstract

In recent years deep neural networks have achieved great success in sentiment classification for English, thanks in part to the availability of copious annotated resources. Unfortunately, most other languages do not enjoy such an abundance of annotated data for sentiment analysis. To tackle this problem, we propose an Adversarial Deep Averaging Network (ADAN) to transfer sentiment knowledge learned from labeled English data to low-resource languages where *only unlabeled* data exists. ADAN is a “Y-shaped” network with two discriminative branches: a *sentiment classifier* and an *adversarial language identification scorer*. Both branches take input from a shared *feature extractor* that aims to learn hidden representations that capture the underlying sentiment of the text and are *invariant* across languages. Experiments on Chinese and Arabic sentiment classification demonstrate that ADAN significantly outperforms several baselines, including a strong pipeline approach that relies on state-of-the-art Machine Translation.

For many other languages, however, only a limited number of sentiment annotations exist. Therefore, previous research in sentiment analysis for low-resource languages focuses on inducing sentiment lexicons (Mohammad et al., 2016b) or training linear classifiers on small domain-specific datasets with hundreds to a few thousand instances (Tan and Zhang, 2008; Lee and Renganathan, 2011). Although some prior work tries to alleviate the scarcity of sentiment annotations by leveraging labeled English data (Wan, 2008, 2009; Lu et al., 2011; Mohammad et al., 2016a), these methods rely on external knowledge such as bilingual lexicons or machine translation (MT) that are expensive to obtain. Some of these papers (Zhou et al., 2016; Wan, 2009) also require translating the entire English training set, which is too demanding if using large amount of English training data.

To aid the creation of sentiment classification systems in such low-resource languages, we propose a framework that leverages the abundant resources for a *source* language (here, English, denoted as SOURCE) to produce sentiment analysis models for a *target* language (TARGET). Our framework is *unsupervised* in the sense that it requires only *unlabeled* text in the target language. In particular, we propose ADAN, an end-to-end adversarial neural network (Goodfellow et al., 2014; Ganin and Lempitsky, 2015). It uses labeled data to train a sentiment classifier for the source language, and simultaneously transfers the learned sentiment analysis knowledge to the target language. Our trained system then directly operates on TARGET texts to predict their sentiment.

We hypothesize that an ideal model for cross-lingual sentiment analysis should learn features that both perform well on sentiment classification for the SOURCE, and are invariant with respect to the shift in language. Therefore, ADAN has two discriminative components: i) a *sentiment clas-*

## 1 Introduction

There has been significant progress on English sentiment analysis in recent years using models based on neural networks (Socher et al., 2013; İrsoy and Cardie, 2014a; Le and Mikolov, 2014; Tai et al., 2015; Iyyer et al., 2015). Most of these, however, rely on a massive amount of labeled training data or fine-grained annotations such as the Stanford Sentiment Treebank (Socher et al., 2013), which provides sentiment annotations for each phrase in the parse tree of every sentence.

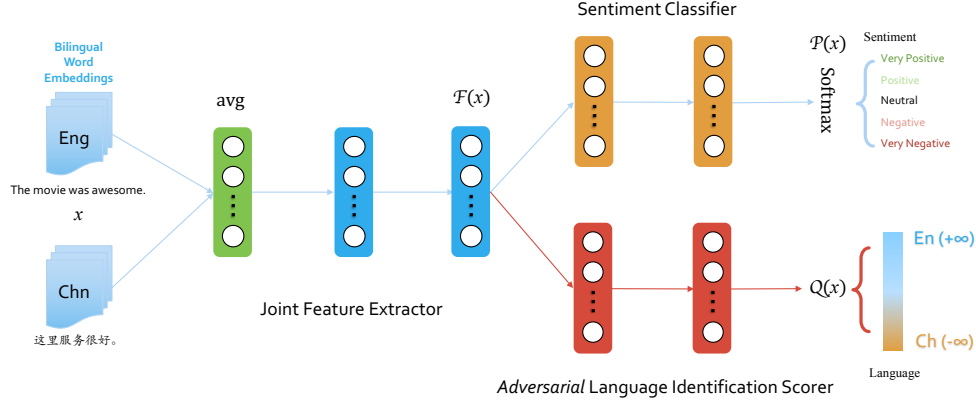


Figure 1: Adversarial Deep Averaging Network with Chinese as the target language. The sentiment classifier  $\mathcal{P}$  and the language scorer  $\mathcal{Q}$  both take input from the feature extractor  $\mathcal{F}$ , and are optimized to excel in their own tasks.  $\mathcal{Q}$  has output width 1, which is deemed as a scalar score indicating how likely a sample is from SOURCE (more in Section 2).  $\mathcal{F}$  aims to learn features that help  $\mathcal{P}$  while **hindering** the adversarial  $\mathcal{Q}$ , in order to learn features helpful for sentiment classification and invariant across languages. Bilingual word embeddings (BWEs) will be discussed in Section 2.1.

sifier  $\mathcal{P}$  for SOURCE; and ii) an adversarial *language identification scorer*  $\mathcal{Q}$  that predicts a scalar indicating whether the input text  $x$  is from the SOURCE (higher score) or the TARGET (lower). The structure of the model is shown in Figure 1. The two classifiers take input from the shared feature extractor  $\mathcal{F}$ , which operates on the average of the bilingual word embeddings (BWEs) for an input text from either SOURCE or TARGET.

While  $\mathcal{P}$  and  $\mathcal{Q}$  each learn to excel in their own task,  $\mathcal{F}$  drives its parameters to extract hidden representations that help the sentiment prediction of  $\mathcal{P}$  and *hamper* the language identification of  $\mathcal{Q}$ . Upon successful training, the joint features (outputs of  $\mathcal{F}$ ) are thus encouraged to be both discriminative for sentiment analysis and invariant across languages. Since ADAN learns language-invariant features by preventing  $\mathcal{Q}$  from identifying the language of a sample,  $\mathcal{Q}$  is hence “**adversarial**”. The intuition is that if  $\mathcal{Q}$  cannot tell the language of a given input sequence using the adversarially trained features, then those features from  $\mathcal{F}$  are effectively language-invariant.

The model is exposed to both SOURCE and TARGET texts during training; SOURCE and TARGET data are passed through the language scorer, while only the labeled SOURCE data pass through the sentiment classifier. The feature extractor and the sentiment classifier are then used for TARGET texts at test time. In this manner, we can train ADAN with labeled SOURCE data and unlabeled TARGET text.

The idea of incorporating an adversary in neural networks has achieved great success in computer vision for image generation (Goodfellow et al.,

2014) and domain adaptation (Ganin and Lempitsky, 2015). However, to our best knowledge, ours is the first to develop an adversarial network for language adaptation, i.e. cross-lingual NLP tasks. In addition, inspired by Arjovsky et al. (2017), we modify the traditional adversarial training method proposed by Ganin and Lempitsky (2015), providing improved performance with smoother training (Sec. 2.2).

We evaluate ADAN using English as SOURCE with both Chinese and Arabic as TARGET, and find that ADAN substantially outperforms i) *train-on-source* cross-lingual approaches trained using labeled SOURCE data; ii) closely related domain adaptation methods, and iii) approaches that employ powerful MT systems. We further investigate the semi-supervised setting, where a small amount of TARGET annotated data exists, and show that ADAN still beats all the baseline systems given the same amount of TARGET supervision. Finally, we provide analysis and visualization of ADAN, shedding light on how it manages to achieve its strong cross-lingual performance. Last but not least, we study a key component in ADAN, the Bilingual Word Embeddings, and demonstrate that ADAN’s performance is robust with respect to the choice of BWEs. Even with random initialized embeddings, ADAN outperforms some of the BWE baselines (Sec. 3.3.3).

## 2 The ADAN Model

### 2.1 Network Architecture

As illustrated in Figure 1, ADAN is a feed-forward network with two branches. There are three main

components in the network, a joint *feature extractor*  $\mathcal{F}$  that maps an input sequence  $x$  to the shared feature space, a *sentiment classifier*  $\mathcal{P}$  that predicts the sentiment label for  $x$  given the feature representation  $\mathcal{F}(x)$ , and a *language scorer*  $\mathcal{Q}$  that also takes  $\mathcal{F}(x)$  but predicts a scalar score indicating whether  $x$  is from SOURCE or TARGET.

An input document is modeled as a sequence of words  $x = w_1, \dots, w_n$ , where each word  $w$  is represented by its word embedding  $v_w$  (Turian et al., 2010). Because the same feature extractor  $\mathcal{F}$  operates on both SOURCE and TARGET sentences, it is favorable if the word representations for both languages align approximately in a shared space. Thus, we employ bilingual word embeddings (BWEs) (Zou et al., 2013; Gouw et al., 2015) to induce distributed word representations that encode semantic relatedness between words across languages, so that similar words are closer in the embedded space regardless of language.

In some prior work, a parallel corpus is required to train the BWEs, making ADAN implicitly “supervised” in the target language. The same can be said for previous work in cross-lingual sentiment classification that requires a sophisticated MT system to link the two languages (Wan, 2009; Zhou et al., 2016). The latter require more direct bilingual supervision in the form of translated SOURCE training data. Thus, this approach is not generally feasible for tasks that require massive amounts of, or evolving, training data. In contrast, ADAN relies on a fixed set of domain-independent BWEs, and no change is necessary when the training data changes. Moreover, even with random initialized embeddings, ADAN can still outperform some baseline methods that use BWEs (see Sec. 3.3.3).

We adopt the Deep Averaging Network (DAN) by Iyyer et al. (2015) for the feature extractor  $\mathcal{F}$ . Although other architectures could be employed, we chose DAN because it is a simple neural network model that yields surprisingly good performance for monolingual sentiment classification. For each document, DAN takes the arithmetic mean of the word vectors as input, and passes it through several fully-connected layers until a softmax for classification. In ADAN,  $\mathcal{F}$  first calculates the average of the word vectors in the input sequence, then passes the average through a feed-forward network with ReLU nonlinearities. The activations of the last layer in  $\mathcal{F}$  are considered the extracted features for the input and are then

passed on to  $\mathcal{P}$  and  $\mathcal{Q}$ . The sentiment classifier  $\mathcal{P}$  and the language scorer  $\mathcal{Q}$  are standard feed-forward networks.  $\mathcal{P}$  has a softmax layer on top for sentiment classification and  $\mathcal{Q}$  ends with a linear layer of output width 1 to assign a language identification score<sup>1</sup>.

## 2.2 Adversarial Training

Consider the distribution of the joint hidden features  $\mathcal{F}$  for both SOURCE and TARGET instances:

$$P_{\mathcal{F}}^{src} \triangleq P(\mathcal{F}(x)|x \in \text{SOURCE})$$

$$P_{\mathcal{F}}^{tgt} \triangleq P(\mathcal{F}(x)|x \in \text{TARGET})$$

As mentioned above, we train  $\mathcal{F}$  to make these two distributions as close as possible to learn language-invariant features for better cross-lingual generalization. Departing from previous research in adversarial training (Ganin and Lempitsky, 2015), in this work we minimize the Wasserstein distance, following Arjovsky et al. (2017). As argued by Arjovsky et al. (2017), existing approaches to training adversarial networks are equivalent to minimizing the Jensen-Shannon distance between two distributions, in our case  $P_{\mathcal{F}}^{src}$  and  $P_{\mathcal{F}}^{tgt}$ . And because Jensen-Shannon suffers from discontinuities, providing less useful gradients for training  $\mathcal{F}$ , Arjovsky et al. (2017) propose instead to minimize the Wasserstein distance and demonstrate its improved stability for hyperparameter selection.

As a result, we too minimize the Wasserstein distance between  $P_{\mathcal{F}}^{src}$  and  $P_{\mathcal{F}}^{tgt}$  according to the Kantorovich-Rubinstein duality (Villani, 2008):

$$W(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt}) = \tag{1}$$

$$\sup_{\|g\|_L \leq 1} \mathbb{E}_{f(x) \sim P_{\mathcal{F}}^{src}} [g(f(x))] - \mathbb{E}_{f(x') \sim P_{\mathcal{F}}^{tgt}} [g(f(x'))]$$

where the supremum (maximum) is taken over the set of all 1-Lipschitz<sup>2</sup> functions  $g$ . In order to (approximately) calculate  $W(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt})$ , we use the language scorer  $\mathcal{Q}$  as the function  $g$  in (1), whose objective is then to seek the supremum in (1) to estimate  $W(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt})$ . To make  $\mathcal{Q}$  a Lipschitz function (up to a constant), the parameters of  $\mathcal{Q}$  are always clipped to a fixed range. Let  $\mathcal{Q}$  be parameterized by  $\theta_q$ , then the objective  $J_q$  of  $\mathcal{Q}$  be-

<sup>1</sup>  $\mathcal{Q}$  simply tries to maximize scores for SOURCE texts and minimize for TARGET, and the scores are not bounded.

<sup>2</sup> A function  $g$  is 1-Lipschitz iff.  $|g(x) - g(y)| \leq |x - y|$  for all  $x$  and  $y$ .

comes:

$$J_q(\theta_f) \equiv \max_{\theta_q} \mathbb{E}_{\mathcal{F}(x) \sim P_{\mathcal{F}}^{src}} [\mathcal{Q}(\mathcal{F}(x))] - \mathbb{E}_{\mathcal{F}(x') \sim P_{\mathcal{F}}^{tgt}} [\mathcal{Q}(\mathcal{F}(x'))] \quad (2)$$

Intuitively,  $\mathcal{Q}$  tries to output higher scores for SOURCE instances and lower scores for TARGET.

For the sentiment classifier  $\mathcal{P}$  parameterized by  $\theta_p$ , we use the traditional cross-entropy loss, denoted as  $L_p(\hat{y}, y)$ , where  $\hat{y}$  and  $y$  are the predicted label distribution and the true label, respectively.  $L_p$  is the negative log-likelihood that  $\mathcal{P}$  predicts the correct sentiment label. We therefore seek the minimum of the following loss function for  $\mathcal{P}$ :

$$J_p(\theta_f) \equiv \min_{\theta_p} \mathbb{E}_{(x,y)} [L_p(\mathcal{P}(\mathcal{F}(x)), y)] \quad (3)$$

Finally, the joint feature extractor  $\mathcal{F}$  parameterized by  $\theta_f$  strives to minimize both the sentiment classifier loss  $J_p$  and  $W(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt}) \approx J_q$ :

$$J_f \equiv \min_{\theta_f} J_p(\theta_f) + \lambda J_q(\theta_f) \quad (4)$$

where  $\lambda$  is a hyper-parameter that balances the two branches  $\mathcal{P}$  and  $\mathcal{Q}$ .

As proved by Arjovsky et al. (2017) and observed in our experiments, minimizing the Wasserstein distance is much more stable w.r.t. hyperparameter selection, saving the hassle of carefully varying  $\lambda$  during training (Ganin and Lempitsky, 2015). In addition, traditional adversarial training methods need to laboriously coordinate the alternating training of the two competing components (Goodfellow et al., 2014) by setting a hyperparameter  $k$ , which indicates the number of iterations one component is trained before training the other. Unfortunately, performance can degrade substantially if  $k$  is not properly set. However, in our case, delicate tuning of  $k$  is no longer necessary since  $W(P_{\mathcal{F}}^{src}, P_{\mathcal{F}}^{tgt})$  is approximated by maximizing (2); thus, training  $\mathcal{Q}$  to optimum using a large  $k$  can provide better performance (but is slower to train). In our experiments,  $\mathcal{F}$  and  $\mathcal{P}$  are trained together and  $\mathcal{Q}$  is trained separately. We use  $\lambda = 0.1$  and  $k = 5$  (train 5  $\mathcal{Q}$  iterations per  $\mathcal{F}$  and  $\mathcal{P}$  iteration), and the performance is stable over a large set of hyperparameters (See Section 3.3.4).

### 3 Experiments and Discussions

To demonstrate the effectiveness of our model, we experiment on Chinese and Arabic sentiment classification, using English as SOURCE for both. For

all data used in experiments, tokenization is done using Stanford CoreNLP (Manning et al., 2014).

#### 3.1 Data

**Labeled English Data.** We use a balanced dataset of 700k Yelp reviews from Zhang et al. (2015) with their sentiment ratings as labels (scale 1-5). We also adopt their train-validation split: 650k reviews for training and 50k form a validation set.

**Labeled Chinese Data.** Since ADAN does not require labeled Chinese data for training, this annotated data is solely used to validate the performance of our model. 10k balanced Chinese hotel reviews from Lin et al. (2015) are used as validation set for model selection and parameter tuning. The results are reported on a separate test set of another 10k hotel reviews. For Chinese, the data are annotated with 5 labels ( $--$ ,  $-$ ,  $0$ ,  $+$ ,  $++$ ).

**Unlabeled Chinese Data.** For the unlabeled TARGET data used in training ADAN, we use another 150k unlabeled Chinese hotel reviews.

**English-Chinese Bilingual Word Embeddings.** For Chinese, we used the pre-trained bilingual word embeddings (BWE) by Zou et al. (2013). Their work provides 50-dimensional embeddings for 100k English words and another set of 100k Chinese words. For more experiments and discussions on BWE, see Section 3.3.3.

**Labeled Arabic Data.** We use the BBN Arabic Sentiment Analysis dataset (Mohammad et al., 2016a) for Arabic sentiment classification. The dataset contains 1200 sentences from social media posts annotated with 3 sentiment labels ( $-$ ,  $0$ ,  $+$ ). The dataset also provides machine translated text to English. Since the label set does not match with the English dataset, we map all the rating 4 and 5 English instances to  $+$  and the rating 1 and 2 instances to  $-$ , while the rating 3 sentences are converted to  $0$ .

**Unlabeled Arabic Data.** For Arabic, no additional unlabeled data is used. We only use the text from the annotated data (without labels) during training.

**English-Arabic Bilingual Word Embeddings.** For Arabic, since no pre-trained BWE is available, we train a 300d BiBOWA BWE (Gouws et al., 2015) on the United Nations corpus (Ziemski et al., 2016).

#### 3.2 Cross-Lingual Sentiment Classification

Our main results are shown in Table 1, which shows very similar trends for Chinese and Ara-

Setting	Approach	Accuracy	
		Chinese	Arabic
Train-on-source-only	Logistic Regression	30.58%	45.83%
	DAN	29.11%	48.00%
Domain Adaptation	mSDA (Chen et al., 2012)	31.44%	48.33%
Machine Translation	Logistic Regression + MT	34.01%	51.67%
	DAN + MT	39.66%	52.50%
Ours	ADAN (CN: 50d, AR:300d)	<b>42.95%<sup>†</sup></b>	<b>55.33%</b>

<sup>†</sup>  $p < 0.001$  under a McNemar test.

Table 1: ADAN performance for Chinese (5-cl) and Arabic (3-cl) sentiment classification without using labeled TARGET data. *All systems* use BWE to map SOURCE and TARGET words into the same space.

bic. Note first that in all of our experiment settings, traditional features like bag of words cannot be directly used since SOURCE and TARGET have completely different vocabularies. Therefore, bilingual word embeddings (BWE) are used as the input representation for *all systems* to map words from both SOURCE and TARGET into the same feature space. In addition, some existing CLSC methods (Wan, 2009; Zhou et al., 2016) need to translate the entire English training set into each target language, which are prohibitive in our setting since our training set has 650k samples.

We start by considering two baselines that train only on the SOURCE language, English, and rely solely on the BWE to classify the TARGET. The first variation uses a standard supervised learning algorithm, Logistic Regression (LR), shown in Row 1 in Table 1. In addition, we evaluate a non-adversarial variation of ADAN, just the DAN portion of our model (Row 2), which is one of the state-of-the-art neural models for sentiment classification. We can see from Table 1 that, in comparison to ADAN (bottom line), BWE by itself does not suffice to transfer knowledge of English sentiment classification to TARGET, and the performance of DAN is poor. On Chinese, even LR performs slightly better, despite that DAN outperforms LR by a large margin on English sentiment classification (not shown in table). This might suggest that fitting tightly on the English data does not necessarily entail good performance on TARGET due to the distributional discrepancy.

We next compare ADAN with domain adaptation baselines, since it can be viewed as a generalization of the cross-lingual task. Nonetheless, domain adaptation methods did not yield satisfactory results for our task. TCA (Pan et al., 2011) did not work since it required quadratic space in terms

of the number of samples (650k). SDA (Glorot et al., 2011) and the subsequent mSDA (Chen et al., 2012) are proven very effective for cross-domain sentiment classification on Amazon reviews. However, as shown in Table 1 (Row 3), mSDA did not perform competitively. We speculate that this is because many domain adaptation models including mSDA were designed for the use of bag-of-words features, which are ill-suited in our task where the two languages have completely different vocabularies. In summary, this suggests that even strong domain adaptation algorithms cannot be used out of the box for our task to get satisfactory results.

Finally, we evaluate ADAN against Machine Translation baselines (Row 4-5) that (1) translate the TARGET text into English and then (2) use the better of the train-on-source-only models for sentiment classification. Previous studies (Banea et al., 2008; Salameh et al., 2015) on sentiment analysis for Arabic and European languages claim this MT approach to be very competitive and can sometimes match the state-of-the-art system trained on that language. For Chinese, where translated text was not provided, we use the commercial Google Translate engine<sup>3</sup>, which is highly engineered, trained on enormous resources, and arguably one of the best MT systems. As shown in Table 1, our ADAN model substantially outperforms the MT baseline on both languages, indicating that our adversarial model can successfully perform cross-lingual sentiment analysis without any annotated data on the target language.

### 3.3 Analysis and Discussions

Since the Arabic dataset is small and may produce noisy results, we chose Chinese as an example for

<sup>3</sup><https://translate.google.com>

our further analysis.

### 3.3.1 Semi-supervised Learning

In practice, it is usually not very difficult to obtain at least a little bit of annotated data. ADAN can be readily adapted to exploit such extra labeled data in the target language, by letting those labeled instances pass through the sentiment classifier  $\mathcal{P}$  as the English samples do during training. We simulate this semi-supervised scenario by adding labeled Chinese reviews for training. We start from adding 100 labeled reviews and keep doubling the number until 12800. As shown in Figure 2, when adding the same number of labeled reviews, ADAN can better utilize the extra supervision and outperform the DAN baseline trained with combined data, as well as the supervised DAN using only labeled Chinese reviews. The margin is naturally decreasing as more supervision is incorporated, but ADAN is still superior when adding 12800 labeled reviews. On the other hand, the DAN with translation baseline seems not able to effectively utilize the added supervision in Chinese, and the performance only starts to show a slightly increasing trend when adding 6400 or more labeled reviews. One possible reason is that when adding into the training data a small number of English reviews translated from the labeled Chinese data, the training signals they produce might be lost in the vast number of English training samples, thus not effectively improving the performance. Another interesting find is that it seems a very small amount of supervision (e.g. 100 labels) could significantly help DAN. However, with the same number of labeled reviews, ADAN still outperforms the DAN baseline.

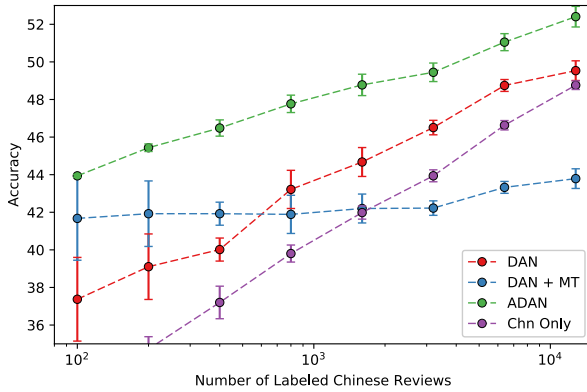


Figure 2: ADAN performance for Chinese in the semi-supervised setting when using various amount of labeled Chinese data.

### 3.3.2 Qualitative Analysis and Visualizations

To qualitatively demonstrate how ADAN bridges the distributional discrepancies between English and Chinese instances, t-SNE (Van der Maaten and Hinton, 2008) visualizations of the activations at various layers are shown in Figure 3. We randomly select 1000 reviews from the Chinese and English validation sets respectively, and plot the t-SNE of the hidden node activations at three locations in our model: the averaging layer, the end of the joint feature extractor, and the last hidden layer in the sentiment classifier before softmax. The train-on-English model is the DAN baseline in Table 1. Note that there is actually only one “branch” in this baseline model, but in order to compare to ADAN, we conceptually treat the first three layers as the feature extractor.

Figure 3a shows that BWE alone does not suffice to bridge the gaps between the distributions of the two languages. Furthermore, we can see in Figure 3b that the distributional discrepancies between Chinese and English are significantly reduced after passing through the joint feature extractor ( $\mathcal{F}$ ), and the learned feature in ADAN brings the distributions in the two languages dramatically closer compared to the monolingually trained baseline. This is measured by the Averaged Hausdorff Distance (Shapiro and Blaschko, 2004; Schütze et al., 2010), which is used to measure the distance between two sets of points. Figure 3 annotates each sub-figure with the AHD between the English and Chinese reviews.

Finally, when looking at the last hidden layer activations in the sentiment classifier of the baseline model (Figure 3c), there are several notable clusters of the red dots (English data) that roughly correspond to the class labels. These English clusters are the areas where the classifier is the most confident in making decisions. However, most Chinese samples are not close to one of those clusters due to the distributional diversion and may thus cause degraded performance in Chinese sentiment classification. On the other hand, the Chinese samples are more in line with the English ones in ADAN, which results in the accuracy boost over the baseline model. In Figure 3, a pair of similar English and Chinese 5-star reviews are highlighted to visualize how the distribution evolves at various points of the network. We can see in 3c that the Chinese review goes close to the “positive English cluster” in ADAN, while in the baseline, it



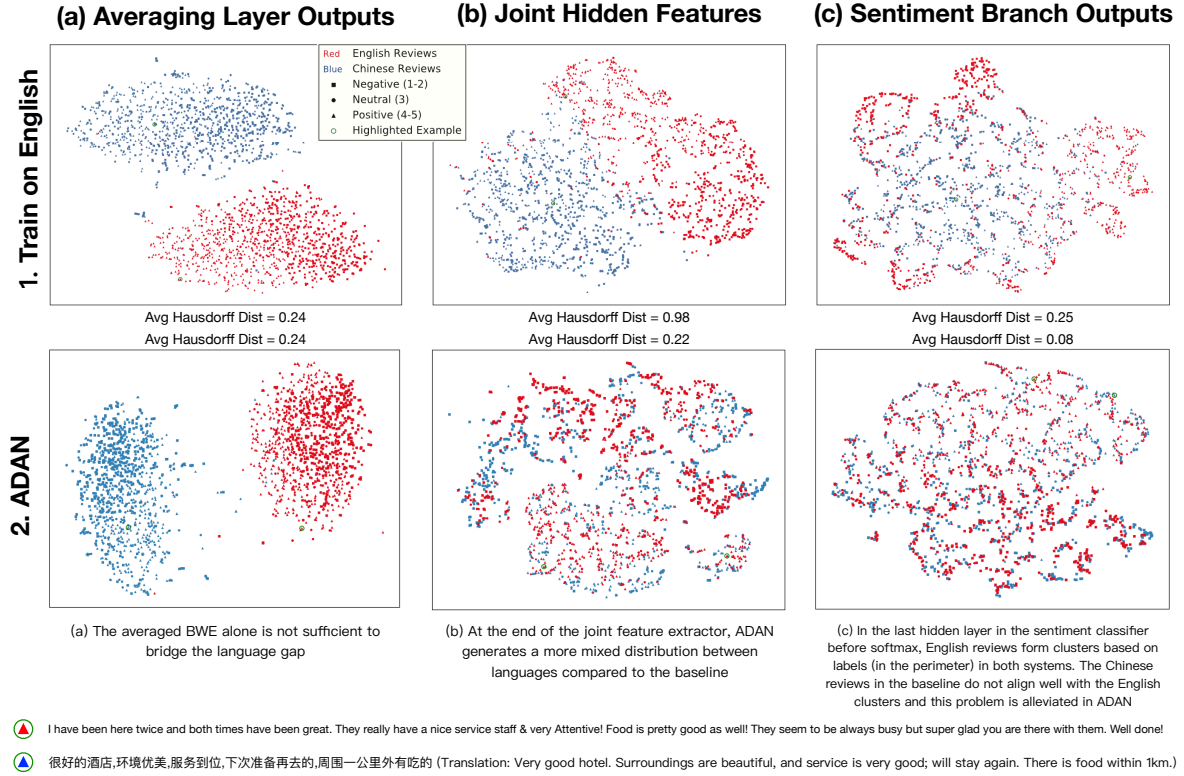


Figure 3: t-SNE Visualizations of activations at various layers for the train-on-source-only baseline model (top) and ADAN (bottom). Better viewed in color and zoom in for more details. The distributions of the two languages are brought much closer in ADAN as they are represented deeper in the network (left to right) measured by the Averaged Hausdorff Distance (discussed later). The green circles are two 5-star example reviews (shown below the figure) that illustrate how the distribution evolves.

stays away from dense English clusters where the sentiment classifier trained on English data are not confident to make predictions.

### 3.3.3 Impact of Bilingual Word Embeddings

In this section we discuss the effect of the bilingual word embeddings. We start by feeding the systems with random initialized WEs, shown in Table 2. ADAN with random WE outperforms the DAN and mSDA baselines using BWE and matches the performance of the LR+MT baseline (Table 1), suggesting ADAN successfully extracts features that could be used for cross-lingual classification tasks similar to BWE without *any* bitext.

With the introduction of BWE, the performance of ADAN is further boosted. Therefore, it seems the quality of BWE plays an important role in cross-lingual classification. To investigate the impact of BWE, we also trained a 100d BilBOWA BWE (Gouws et al., 2015) using the UN parallel corpus for Chinese. All systems achieve slightly lower performance compared to the pre-trained BWE, yet ADAN still outperforms other baseline methods (Table 2), demonstrating that ADAN’s effectiveness is relatively robust with respect to the

choice of BWE. For the reason why all systems show inferior results with BilBOWA, we conjecture that BilBOWA may have slightly reduced quality since it does not require word alignments during training as by Zou et al. (2013). By only training on sentence-aligned corpus, BilBOWA requires less resource and is much faster to train, potentially at the expense of quality.

Model	Random	BilBOWA	Pre-trained
DAN	21.66%	28.75%	29.11%
DAN+MT	37.78%	38.17%	39.66%
ADAN	34.44%	40.51%	42.95%

Table 2: Model performance for various (B)WE choices for Chinese.

### 3.3.4 ADAN Hyperparameter Stability

In this section, we show that the training of ADAN is stable over a large set of hyperparameters, and provide improved performance compared to traditional adversarial training method by Ganin and Lempitsky (2015).

We implemented a variant of ADAN similar to the adversarial domain adaptation network

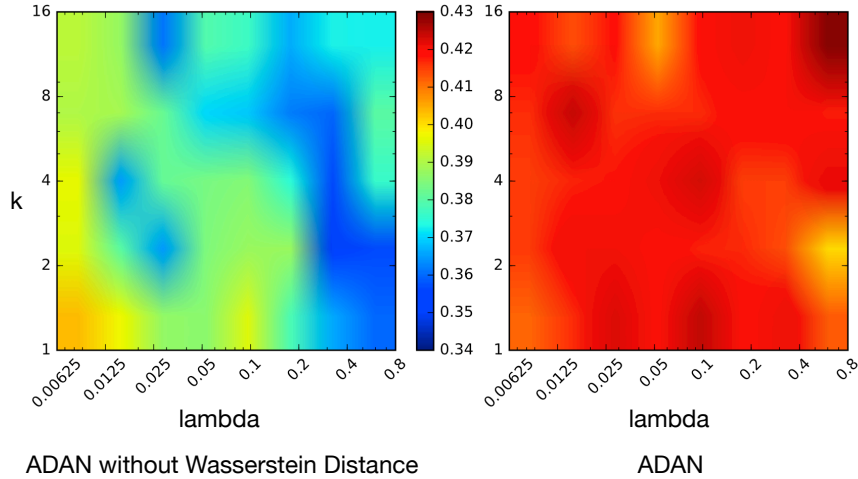


Figure 4: A grid search on  $k$  and  $\lambda$  for ADAN (right) and the Ganin and Lempitsky (2015) variant (left). Numbers indicate the accuracy on the Chinese development set.

by Ganin and Lempitsky (2015). In particular,  $Q$  is now a binary classifier with a softmax layer on top which classifies if an input text  $x$  is from SOURCE or TARGET, given its hidden features  $\mathcal{F}(x)$ . For training,  $Q$  is connected to  $\mathcal{F}$  with a GradientReversalLayer (Ganin and Lempitsky, 2015) in between, which preserves the input during the a forward pass but multiplies the gradients by  $-\lambda$  during a backward pass.  $\lambda$  is a hyperparameter similar to that in ADAN that balances the effects  $\mathcal{P}$  and  $Q$  have on  $\mathcal{F}$  respectively. This way, the entire network can be trained in its entirety using standard backpropagation. In addition, as mentioned in Section 2.2, the training of  $\mathcal{F}$  and  $Q$  might not be fully in sync, and efforts need to be made to coordinate the adversarial training. This is achieved by setting  $\lambda$  to a non-zero value only once out of  $k$  batches. Here,  $k$  is again a hyperparameter similar to that in ADAN which corresponds the training of  $\mathcal{F}$  and  $Q$ . When  $\lambda = 0$ , the gradients from  $Q$  will not be back-propagated to  $\mathcal{F}$ . This allows  $Q$  more iterations to adapt to  $\mathcal{F}$  before  $\mathcal{F}$  makes another adversarial update.

To verify the superiority of ADAN, we conduct a grid search over the two hyperparameters:  $k$  (number of  $Q$  iterations per  $\mathcal{F}$  iteration) and  $\lambda$  (the balance factor between  $\mathcal{P}$  and  $Q$ ). We experiment with  $k \in \{1, 2, 4, 8, 16\}$ , and  $\lambda \in \{0.00625, 0.0125, 0.025, 0.05, 0.1, 0.2, 0.4, 0.8\}$ . Figure 4 reports the accuracy on the Chinese development set of both ADAN variants. In Figure 4, it can be clearly seen that ADAN achieves higher accuracy while being much more stable

than the Ganin and Lempitsky (2015) variant, suggesting that ADAN overcomes the well-known problem that adversarial training is sensitive to hyperparameter tuning.

### 3.4 Implementation Details

For all our experiments on both languages, the feature extractor  $\mathcal{F}$  has three fully-connected hidden layers with ReLU non-linearities, while both  $\mathcal{P}$  and  $Q$  have two. All hidden layers contains 900 hidden units. This choice is more or less ad-hoc, and the performance could potentially be improved with more careful model selection. Batch Normalization (Ioffe and Szegedy, 2015) is used in each hidden layer in  $\mathcal{P}$  and  $Q$ .  $\mathcal{F}$  does not use BN.  $\mathcal{F}$  and  $\mathcal{P}$  are optimized together by Adam (Kingma and Ba, 2015) with a learning rate of 0.05 for Chinese and 0.01 for Arabic experiments.  $Q$  is trained with Adam with learning rate of 0.00005. The weights of  $Q$  are clipped to  $[-0.01, 0.01]$ . ADAN is implemented in Torch7 (Collobert et al., 2011). We train ADAN for 30 epochs and use early stopping to select the best model on the validation set.

## 4 Related Work

**Cross-lingual Sentiment Analysis** is motivated by the lack of high-quality labeled data in many non-English languages (Mihalcea et al., 2007; Banea et al., 2008, 2010). For Chinese and Arabic in particular, there are several representative works (Wan, 2008, 2009; He et al., 2010; Lu et al., 2011; Mohammad et al., 2016a). Our work is



comparable to these papers in objective but very different in method. The work by Wan uses machine translation to directly convert English training data to Chinese; this is one of our baselines. Lu et al. (2011) instead uses labeled data from both languages to improve the performance on both.

**Domain Adaptation** tries to learn effective classifiers for which the training and test samples are from different underlying distributions (Blitzer et al., 2007; Pan et al., 2011; Glorot et al., 2011; Chen et al., 2012; Liu et al., 2015). This can be thought of as a generalization of cross-lingual text classification. However, one main difference is that, when applied to text classification tasks such as sentiment analysis, most work assumes a common feature space such as bag of words, which is not available in the cross-lingual setting. See Section 3.2 for experiments on this. In addition, most works in domain adaptation evaluate on adapting product reviews across domains (e.g. books to electronics), where the divergence in distribution is less significant than that between two languages.

**Adversarial Networks** have enjoyed much success in computer vision (Goodfellow et al., 2014; Ganin and Lempitsky, 2015), but to our best knowledge, have not yet achieved comparable success in NLP. We are the first to apply adversarial training to cross-lingual NLP tasks. A series of work in image generation has used architectures similar to ours, by pitting a neural image generator against a discriminator that learns to classify real versus generated images (Goodfellow et al., 2014; Denton et al., 2015). More relevant to this work, adversarial architectures have produced the state-of-the-art in unsupervised domain adaptation for image object recognition: Ganin and Lempitsky (2015) train with many labeled source images and unlabeled target images, similar to our setup. In addition, some recent work (Arjovsky et al., 2017; Gulrajani et al., 2017) propose improved methods for training Generative Adversarial Nets.

## 5 Conclusion and Future Work

In this work, we presented ADAN, an adversarial deep averaging network for cross-lingual sentiment classification, which, for the first time, applies adversarial training to cross-lingual NLP. ADAN leverages the abundant resources on English to help sentiment analysis on other languages where little or no annotated data exist. We validate our hypothesis by empirical experiments on Chi-

nese and Arabic sentiment classification, where we have labeled English data and only *unlabeled* data in the target language. Experiments show that ADAN outperforms several baselines including domain adaptation models and a highly competitive MT baseline. We further show that even without *any* bilingual resources, ADAN trained with random initialized embeddings can still achieve meaningful cross-lingual performance. In addition, we show that in the presence of labeled data in the target language, ADAN can naturally incorporate this additional supervision and yields even more competitive results.

For future work, we plan to apply our adversarial training framework to other NLP adaptation tasks, where explicit MLE training is not feasible due to the lack of direct supervision. For instance, our framework is not limited to text classification tasks, and can be extended to phrase level opinion mining (İrsoy and Cardie, 2014b) by extracting phrase-level opinion expressions from sentences using deep recurrent neural networks. Our framework can be applied to these phrase-level models for languages where labeled data might not exist. In another direction, our adversarial framework for cross-lingual text categorization can be used in conjunction with not only DAN, but also many other neural models such as LSTM, etc.

## References

- M. Arjovsky, S. Chintala, and L. Bottou. 2017. Wasserstein GAN. *ArXiv e-prints* <https://arxiv.org/abs/1701.07875>.
- Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, pages 28–36. <http://aclweb.org/anthology/C10-1004>.
- Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 127–135. <http://aclweb.org/anthology/D08-1014>.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. As-

- sociation for Computational Linguistics, pages 440–447. <http://aclweb.org/anthology/P07-1056>.
- Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, Omnipress, New York, NY, USA, ICML '12, pages 767–774. <http://icml.cc/2012/papers/416.pdf>.
- Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. 2011. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*. [http://cs.nyu.edu/~koray/files/2011\\_torch7\\_nipsw.pdf](http://cs.nyu.edu/~koray/files/2011_torch7_nipsw.pdf).
- Emily L Denton, Soumith Chintala, arthur szlam, and Rob Fergus. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., pages 1486–1494. <http://papers.nips.cc/paper/5773-deep-generative-image-models-using-a-laplacian-pyramid-of-adversarial-networks.pdf>.
- Yaroslav Ganin and Victor Lempitsky. 2015. Un-supervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning. JMLR Workshop and Conference Proceedings*. <http://jmlr.org/proceedings/papers/v37/ganin15.pdf>.
- Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. ACM, New York, NY, USA, ICML '11, pages 513–520. [http://www.icml-2011.org/papers/342\\_icmlpaper.pdf](http://www.icml-2011.org/papers/342_icmlpaper.pdf).
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 2672–2680. <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>.
- Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*. <http://jmlr.org/proceedings/papers/v37/gouws15.pdf>.
- I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. 2017. Improved Training of Wasserstein GANs. *ArXiv e-prints*.
- Yulan He, Harith Alani, and Deyu Zhou. 2010. Exploring english lexicon knowledge for chinese sentiment analysis. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*. <http://aclweb.org/anthology/W10-4116>.
- Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*. <http://jmlr.org/proceedings/papers/v37/ioffe15.pdf>.
- Ozan Irsoy and Claire Cardie. 2014a. Deep recursive neural networks for compositionality in language. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 2096–2104. <http://papers.nips.cc/paper/5551-deep-recursive-neural-networks-for-compositionality-in-language.pdf>.
- Ozan Irsoy and Claire Cardie. 2014b. Opinion mining with deep recurrent neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 720–728. <http://aclweb.org/anthology/D14-1080>.
- Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1681–1691. <https://doi.org/10.3115/v1/P15-1162>.
- Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*. <https://arxiv.org/abs/1412.6980>.
- Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*. <http://jmlr.org/proceedings/papers/v32/le14.html>.
- Huey Yee Lee and Hemnaath Renganathan. 2011. Chinese sentiment analysis using maximum entropy. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*. Asian Federation of Natural Language Processing, pages 89–93. <http://aclweb.org/anthology/W11-3713>.
- Yiyou Lin, Hang Lei, Jia Wu, and Xiaoyu Li. 2015. An empirical study on sentiment classification of chinese review using word embedding. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*. pages 258–266. <http://aclweb.org/anthology/Y15-2030>.

- Biao Liu, Minlie Huang, Jiashen Sun, and Xuan Zhu. 2015. Incorporating domain and sentiment supervision in representation learning for domain adaptation. In *International Joint Conference on Artificial Intelligence*. <http://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/10722>.
- Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 320–330. <http://aclweb.org/anthology/P11-1033>.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. <http://www.aclweb.org/anthology/P/P14/P14-5010>.
- Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, pages 976–983. <http://aclweb.org/anthology/P07-1123>.
- Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016a. How translation alters sentiment. *Journal of Artificial Intelligence Research* 55(1):95–130. <http://dl.acm.org/citation.cfm?id=3013558.3013562>.
- Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016b. Sentiment lexicons for arabic social media. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*. <http://www.lrec-conf.org/proceedings/lrec2016/pdf/234.Paper.pdf>.
- S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210. <https://doi.org/10.1109/TNN.2010.2091281>.
- Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 767–777. <https://doi.org/10.3115/v1/N15-1078>.
- Oliver Schütze, Xavier Esquivel, Adriana Lara, and Carlos A Coello Coello. 2010. Measuring the averaged hausdorff distance to the pareto front of a multi-objective optimization problem. Technical report, Technical Report TR-OS-2010-02, CINVESTAV. [http://delta.cs.cinvestav.mx/schuetze/technical\\_reports/TR-OS-2010-02.pdf](http://delta.cs.cinvestav.mx/schuetze/technical_reports/TR-OS-2010-02.pdf).
- Michael D Shapiro and Matthew B Blaschko. 2004. On hausdorff distance measures. Technical report, Technical Report UM-CS-2004-071. <https://web.cs.umass.edu/publication/docs/2004/UM-CS-2004-071.pdf>.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, D. Christopher Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1631–1642. <http://aclweb.org/anthology/D13-1170>.
- Sheng Kai Tai, Richard Socher, and D. Christopher Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1556–1566. <https://doi.org/10.3115/v1/P15-1150>.
- Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. *Expert Syst. Appl.* 34(4):2622–2629. <https://doi.org/10.1016/j.eswa.2007.05.028>.
- Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 384–394. <http://aclweb.org/anthology/P10-1040>.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* <http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>.
- Cédric Villani. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.
- Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 553–561. <http://aclweb.org/anthology/D08-1058>.
- Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*

of the AFNLP: Volume 1 - Volume 1. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '09, pages 235–243. <http://dl.acm.org/citation.cfm?id=1687878.1687913>.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems* 28, Curran Associates, Inc., pages 649–657. <http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf>.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1403–1412. <https://doi.org/10.18653/v1/P16-1133>.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus. In *Language Resources and Evaluation (LREC16)*.

Y. Will Zou, Richard Socher, Daniel Cer, and D. Christopher Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1393–1398. <http://aclweb.org/anthology/D13-1141>.